VIGC: Visual Instruction Generation and Correction

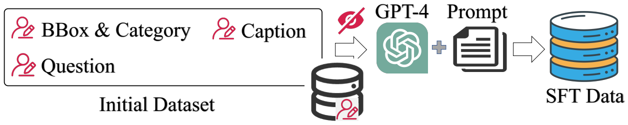
Bin Wang[*1], **Fan Wu**[*1], **Xiao Han**[*1], **Jiahui Peng**[*1], **Huaping Zhong**[*2],
Pan Zhang[1], **Xiaoyi Dong**[1,3], **Weijia Li**[4], **Wei Li**[1], **Jiaqi Wang**[1], **Conghui He**[†1],
[1]Shanghai Artificial Intelligence Laboratory,
[2]SenseTime Research,
[3]The Chinese University of Hong Kong,
[4]Sun Yat-sen University
{wangbin,wufan,hanxiao,pengjiahui,zhangpan,dongxiaoyi,liwei,wangjiaqi,heconghui}@pjlab.org.cn
zhonghuaping@sensetime.com, liweij29@mail.sysu.edu.cn

Abstract

The integration of visual encoders and large language models (LLMs) has driven recent progress in multimodal large language models (MLLMs). However, the scarcity of high-quality instruction-tuning data for vision-language tasks remains a challenge. The current leading paradigm, such as LLaVA, relies on language-only GPT-4 to generate data, which requires pre-annotated image captions and detection bounding boxes, suffering from understanding image details. A practical solution to this problem would be to utilize the available multimodal large language models to generate instruction data for vision-language tasks. However, it's worth noting that the currently accessible MLLMs are not as powerful as their LLM counterparts, as they tend to produce inadequate responses and generate false information. As a solution for addressing the current issue, this paper proposes the Visual Instruction Generation and Correction (VIGC) framework that enables multimodal large language models to generate instruction-tuning data and progressively enhance its quality on-the-fly. Specifically, Visual Instruction Generation (VIG) guides the vision-language model to generate diverse instruction-tuning data. To ensure generation quality, Visual Instruction Correction (VIC) adopts an iterative update mechanism to correct any inaccuracies in data produced by VIG, effectively reducing the risk of hallucination. Leveraging the diverse, high-quality data generated by VIGC, we finetune mainstream models and validate data quality based on various evaluations. Experimental results demonstrate that VIGC not only compensates for the shortcomings of language-only data generation methods, but also effectively enhances the benchmark performance. The models, datasets, and code are available at https://opendatalab.github.io/VIGC.

Introduction

Over the past year, significant advancements have emerged in language models, particularly with instruction tuning in Large Language Models (LLMs). This technology enables models to perform complex tasks in a zero-shot manner (OpenAI 2023a,b). The fusion of visual encoders with LLMs (Touvron et al. 2023; Chiang et al. 2023) has led to substantial strides in the field of multimodal LLMs, resulting in the creation of frameworks such as BLIP-2 (Li et al. 2023b), MiniGPT-4 (Zhu et al. 2023b), LLaVA (Liu et al. 2023b), InstructBLIP (Dai et al. 2023) and InternLM-XComposer (Zhang et al. 2023). These frameworks have propelled the rapid evolution of multimodal tasks, exhibiting impressive capabilities in image-text dialogue.



(a) Text-only GPT-4

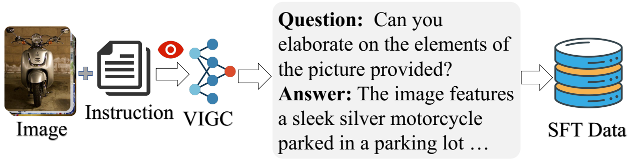


Image
Instruction
VIGC
Question: Can you elaborate on the elements of the picture provided?
Answer: The image features a sleek silver motorcycle parked in a parking lot …
SFT Data

(b) The Proposed VIGC

Figure 1: Comparison between the language-only GPT-4 approach and the proposed method, highlighting two key limitations of the former: (1) The necessity for extensive human annotation, and (2) The inability to process visual data, resulting in a loss of detailed information.

Traditional multimodal models follow a two-stage training process. The initial stage involves training the model with image-text pairs to enhance feature alignment between the two modalities. The subsequent stage utilizes high-quality multimodal instruction tuning data to augment the model's ability to follow instructions, thereby improving its response to user inquiries. However, compared to a large amount of available multimodal pre-training data (Schuhmann et al. 2022; Sharma et al. 2018; Changpinyo et al. 2021; He et al. 2023), acquiring high-quality instruction tuning data is relatively more challenging. Current high-quality multimodal fine-tuning data (Liu et al. 2023b; Li et al. 2023a) is primarily generated based on language-only

---

[*] These authors contributed equally.

[†] Corresponding author.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

GPT-4 (OpenAI 2023b) as illustrated in Figure 1-(a). This approach necessitates costly manual pre-annotation and restricts the design of questions and generated responses to existing annotated information. Consequently, if the question posed is not within this annotated information, GPT-4 is unable to respond. This method also loses the detailed information in the image for answerable questions.

To address this issue, researchers have started to consider generating data with Vision-Language Models (VLMs) (Zhu et al. 2023a; You et al. 2023) as VLMs have seen a vast amount of image-text pairs during the pre-training phase and inherently contain a wealth of knowledge. Currently, the accessible MLLMs are less powerful than their LLM counterparts. They often produce inadequate responses and generate false information, e.g., hallucination. However, existing approaches attempt to generate data using VLMs without considering how to ensure the quality of the generated data or validate it experimentally.

In contrast to the aforementioned methods, we propose Visual Instruction Generation and Correction, a new method for high-quality instruction data generation. This method, based on existing visual-language models, guides the model to generate diverse visual-language question-answer pairs on new images through the fine-tuning of initial instruction data. The ability to generate diverse data is derived from the fact that both the visual encoder and the large language model have been fine-tuned on extensive datasets, encompassing rich image understanding and logical language capabilities. However, we found that data generated directly from provided instructions suffer from severe hallucination issues, which is a common problem plaguing large multimodal models (Peng et al. 2023b; Liu et al. 2023a; Zhao et al. 2023; Huang et al. 2023). Fortunately, our visual instruction correction module can significantly reduce model hallucination phenomena through iterative updates. The primary contributions of this study include:

- We introduce Visual Instruction Generation and Correction (VIGC), a framework capable of autonomously generating high-quality image-text instruction fine-tuning datasets. The VIGC framework consists of two submodules: Visual Instruct Generation (VIG) and Visual Instruct Correction (VIC). Specifically, the VIG generates initial visual question-answer pairs, while VIC mitigates model hallucination and obtains high-quality data through an Iterative Q-Former (IQF) update strategy.
- We release a series of datasets[1] (He et al. 2022) generated using VIGC, including 36,781 VIGC-LLaVA-COCO and approximately 1.8 million VIGC-LLaVA-Objects365, for research on large multimodal models. To the best of our knowledge, this is the first-ever multimodal instruction fine-tuning dataset autonomously generated by a MLLM.
- We have conducted extensive experiments on the generated data. When trained in conjunction with the VIGC-generated data, the performance of the LLaVA-7B model significantly improved, even surpassing that of the LLaVA-13B model. Furthermore, on mainstream multimodal evaluation datasets such as MMBench, OKVQA, and A-OKVQA, models trained with the VIGC data uniformly demonstrated performance enhancements.

Related Work

Instruction-following LLMs

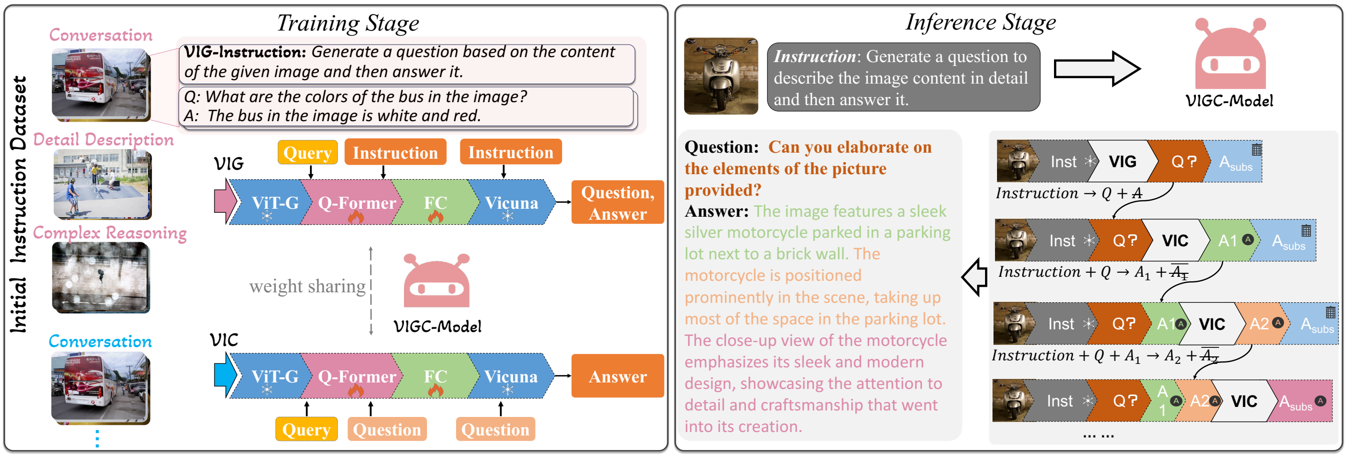
The domain of Natural Language Processing (NLP) has been significantly shaped by the advent and evolution of large language models (LLMs), including but not limited to GPT-3 (Brown et al. 2020), PaLM (Chowdhery et al. 2022), T5 (Raffel et al. 2020), and OPT (Zhang et al. 2022). These models, equipped with extensive training data and sophisticated optimization techniques, have demonstrated remarkable performance across various tasks. However, a notable challenge persists in their ability to effectively follow instructions, often leading to suboptimal results in diverse real-world applications. Efforts to address this issue have led to the introduction of various instruction fine-tuning datasets. Enhanced models, such as InstructGPT (Ouyang et al. 2022), ChatGPT (OpenAI 2023a), FLAN-T5 (Chung et al. 2022), FLAN-PaLM (Chung et al. 2022), and OPT-IML (Iyer et al. 2022), have been developed to improve upon zero-shot and few-shot learning capabilities, primarily by learning to map instructions to the corresponding expected outputs. Despite these advancements, the generation of instruction datasets frequently relies on pre-existing NLP tasks, which curtails their generalizability. To augment the quality and diversity of instructions, Wang et al. (Wang et al. 2022) introduce SELF-INSTRUCT, a methodology that employs generated instruction data to enhance the performance of LLMs. While these methods have made significant strides in augmenting the instruction-following capabilities of language models, they exhibit a standard limitation in that they cannot be directly generalized to multimodal data.

Multi-modal Instruction Tunning

Compared to creating language instruction fine-tuning datasets, constructing multimodal instruction fine-tuning datasets requires a thorough understanding of image content and the development of the corresponding text. MiniGPT-4 utilizes a feature-aligned model to interpret the CC dataset (Sharma et al. 2018; Changpinyo et al. 2021), employs ChatGPT for initial filtering, and ultimately curates 3,500 image-text pairs for model refinement. However, this methodology encounters restrictions in terms of instruction diversity and volume. In contrast, LLaVA proposes an innovative approach based on a language-only GPT-4 (OpenAI 2023b) to generate multimodal instruction data from information that includes caption descriptions and target data. While this approach generates high-quality data, it demands manual annotation of each caption description, target information, and question, which inherently limits scalability. To extend data across a more comprehensive array of tasks, InstructBLIP pioneers an Instruction template construction methodology, converting 26 unique datasets into instruction fine-tuning data and achieving impressive results across several tasks.

---

[1] https://opendatalab.com/OpenDataLab/VIGC-InstData



VIG-Instruction: *Generate a question based on the content of the given image and then answer it.*

Q: What are the colors of the bus in the image?
A: The bus in the image is white and red.

Instruction: Generate a question to describe the image content in detail and then answer it.

Question: Can you elaborate on the elements of the picture provided?
Answer: The image features a sleek silver motorcycle parked in a parking lot next to a brick wall. The motorcycle is positioned prominently in the scene, taking up most of the space in the parking lot. The close-up view of the motorcycle emphasizes its sleek and modern design, showcasing the attention to detail and craftsmanship that went into its creation.

Figure 2: The proposed Visual Instruction Generation and Correction (VIGC) framework. The left panel illustrates the VIGC training process: Instruction fine-tuning data is fed into the VIG and VIC sub-modules. VIG aims to generate image-related question-answer pairs, while VIC refines the VIG-produced answers for precision. The right panel depicts the inference phase, where VIGC takes an arbitrary image as input, generates initial answers, and then refines them to construct high-quality data.

Concurrently, MIMIC (Li et al. 2023a) assembles larger-scale instruction fine-tuning datasets.

Nevertheless, all these datasets require human intervention in the form of annotations, and their diversity is inherently limited by the existing data. By contrast, our study aims to propose a self-guided, model-driven instruction fine-tuning data generation method, which is capable of creating high-quality fine-tuning data suitable for any novel image.

Visual Question Generation

Visual Question Generation (VQG) aims to generate relevant questions based on provided images, which poses considerable challenges due to the need for diversity, naturalness, and engagement. Mostafazadeh *et al.* (Mostafazadeh et al. 2016) propose the task of Visual Question Generation (VQG) and attempt to establish a foundational VQG framework, employing both retrieval-based and generative methodologies. iQAN (Li et al. 2018) later proposed a unified, reversible network addressing both VQA and VQG tasks, enabling both answer retrieval and question generation from images. Guiding models like Guiding Visual Question Generation (Vedd et al. 2021) have also contributed significantly to the field.

This paper proposes the Visual Instruction Generation and Correction network, a model that generates image-related content, similar to VQG. Unlike the existing work, our method introduces an additional layer of complexity by developing diverse questions and providing appropriate answers based on different requirement categories. Leveraging the vast knowledge of large language models, our model's output outperforms traditional VQG tasks, which are usually limited by their training sample size.

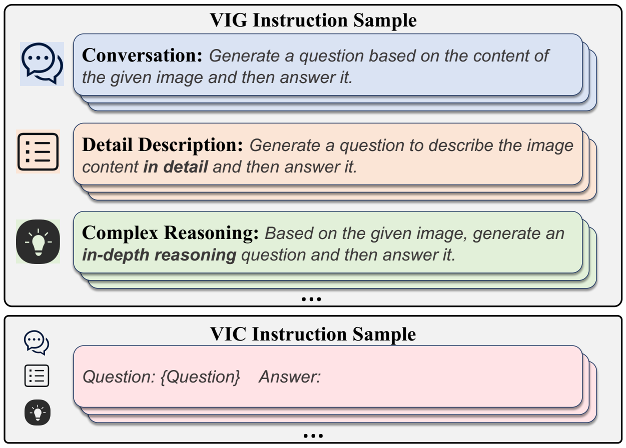
Methods

This paper concentrates on leveraging the power of existing vision-language models to generate multimodal instruction following data autonomously. The proposed approach facilitates the creation of robust and diverse fine-tuning datasets, eliminating the requirement for intensive manual intervention. However, utilizing existing multimodal models to achieve this objective presents substantial challenges. To mitigate these, we introduce a self-instructing framework named VIGC. Guided by existing fine-tuning data, this framework can generate higher quality and more diverse new data, as depicted in Figure 2.

Initial Instruction Construction

In contrast to language instructions, which can be effortlessly generated by standalone language models (Peng et al. 2023a; Wang et al. 2022), the construction of visual-text multimodal instructions requires a detailed understanding of visual content, as well as the ability to pose relevant questions and provide correct answers based on the actual content of the images. Nevertheless, existing multimodal models are deficient in their capacity to directly generate visual-language instruction data. To overcome this limitation, we exploit readily available instruction fine-tuning data and formulate additional instruction templates, thereby facilitating the automatic generation of instruction data.

Our proposed method is universally applicable to generating various types of image-text multimodal instruction fine-tuning data. To elucidate our approach, we exemplify it using the generation of LLaVA-style data instructions. Specifically, we construct instruction templates encompassing dialogue, detailed description, and complex reasoning, following the categorization of instruction fine-tuning data types as delineated in LLaVA. Figure 3 presents instances of these three types of instruction templates, which are essentially uncomplicated, principally requesting, *"generate T-type question-answer pairs predicated on the image content."* Theoretically, if a model can comply with these instruction descriptions following training, it should be profi-



VIG Instruction Sample

Conversation: *Generate a question based on the content of the given image and then answer it.*

Detail Description: *Generate a question to describe the image content **in detail** and then answer it.*

Complex Reasoning: *Based on the given image, generate an **in-depth reasoning** question and then answer it.*

…

VIC Instruction Sample

Question: {Question}   Answer:

…

Figure 3: Template examples corresponding to instruction tuning in VIG and VIC submodules.

cient in generating question-answer pairs.

With the instruction templates and existing visual instruction-tuning data (i.e., Question-Answer pairs in LLaVA), we construct a comprehensive VIG instruction-tuning dataset as follows:

$$T_{VIG} = (X_i, I_t, Q_i^t, A_i^t)^{N_t} \tag{1}$$

where $i \in \{1, 2, ..., N_t\}$, N_t denotes the instruction type, such conversation, detailed description, etc. X_i represents an RGB image, I_i represents an instruction corresponding to a specific type t, Q_i^t is a question related to the image X_i under the context of instruction I_t, and A_i^t is the answer to the question Q_i^t. Our objective is to leverage this dataset for the training of models that, given a specific instruction I_t, can generate corresponding question-answer pairs for a given image, following the designated instruction type. Figure 2 provides illustrations of the initial instruction dataset.

Distinguished from the VIG, the VIC instruction employs an image and a query as input for its fine-tuning process, with the objective of generating precise responses. The dataset for the VIC instruction is presented below:

$$T_{VIC} = (X_i, Q_i^t, A_i^t)^{N_t} \tag{2}$$

Visual Instruction Generation

In alignment with current popular multimodal models such as MiniGPT-4 (Zhu et al. 2023b) and InstructBLIP (Dai et al. 2023), the architecture of the proposed VIGC can be dissected into four primary components: the visual encoder (ViT) (Fang et al. 2023), the large language model (Vicuna) (Chiang et al. 2023), the Q-Former (Li et al. 2023b) for visual feature extraction, and the Fully-Connected (FC) projection for reconciling visual-language features. Functionally, the model can be further segmented into two distinctive sub-modules: the Visual Instruction Generation (VIG) module and the Visual Instruction Correction (VIC) module. It is imperative to underscore that these two sub-modules share network parameters, the primary differentiator being the data type employed for training.

The principal objective of the VIG module is to autonomously produce relevant visual question-answer pairs that correspond to a specific instructional command for any given image. Figure 2 illustrates the process that the VIG module follows in the training phase. In the training phase, the VIG module stochastically selects an image, which is subsequently processed via a visual encoder. It generates a set of fixed visual feature embeddings. The Q-Former module, purposefully designed to be aware of instructional information, further refines these visual features. At this stage, the model employs learnable visual queries that perform self-attention operations in conjunction with the instruction. This operation is followed by a cross-attention phase with visual embeddings. This mechanism impels the visual features to concentrate on the instructional information, thereby augmenting their relevance and precision within the context of the assigned task. Following the cross-attention phase, the refined features are channeled through an FC mapping layer, a crucial step that aligns visual features with their linguistic counterparts, thereby ensuring a seamless integration of visual and language features. Subsequently, the instruction-aligned features are ingested by the language model. This process guides the model to generate the predicted results. Specifically, the objective in this context is to generate visual questions and answers that are intrinsically linked to the content of image X_i, the nature of which is determined by the instruction. We utilize the original auto-regressive loss function inherent to the large language model. This methodology guides the model in generating sentences that align with the question-answer pairs provided in the training set.

Visual Instruction Correction

In the exploration conducted for this study, we discovered that existing multimodal models (Liu et al. 2023b), (Dai et al. 2023), much like language models (Radford et al. 2018, 2019; Brown et al. 2020; OpenAI 2023b,a), often exhibit hallucination issues. This hallucination phenomenon is also present in the data generated by the VIG, especially in instances of extensive descriptions. We attribute this to the tendency of multimodal models to progressively rely on the current answer text during the answer generation phase, thereby gradually neglecting the image information and consequently leading to the description of targets not present in the image. To eliminate the hallucination phenomenon in generated data and ensure that downstream tasks based on this data are not contaminated, we specifically introduce an instruction correction module to update the answers and reduce the occurrence of hallucinations.

To effectively utilize the VIC, specific actions need to be undertaken during both the model training and inference stages:

During the training phase: The goal of the VIG phase is to generate corresponding visual question-answer pairs given an instruction. Conversely, the objective of the VIC training phase is to supply the model with a Question, thereby directing the model to focus on extracting features pertinent to the input question/text during the Q-Former feature extraction process. These features lay the groundwork for subsequent answers.

During the inference phase: After training the model using the aforementioned VIC method, it can take the questions from the question-answer pairs generated by the VIG as input and regenerate answers. Since the model places greater emphasis on the question when formulating responses, the generated results are typically more accurate. Furthermore, we iterate this Q-Former feature updating process, termed as Iterativate-Q-Former (IQF), as illustrated in the VIGC inference phase in Figure 2. Before deploying the VIC module, we initially generate the initial question (Q) and answer (A) using the VIG. In the first iteration, we use the Instruction and Question as inputs to output answers A_1 and $\bar{A}_1$, where A_1 represents the first sentence of the answer and $\bar{A}_1$ signifies all content following the first sentence. In the second iteration, we input the Instruction, Question, and the answer A_1 from the previous step to predict A_2, and this process continues iteratively until a termination symbol is encountered. The efficacy of this iterative approach is primarily due to the continual updating of visual features with the most recent textual information, making subsequent results more accurate. However, it should be noted that while this method is highly beneficial for providing detailed descriptions of image content, its effectiveness for dialogue tasks and inference tasks is relatively limited. This is because dialogue tasks usually consist of single sentences, and the subsequent content in inference tasks does not heavily depend on image information.

Experiments

Datasets

Training Data. We trained the VIGC network using two types of visual-language instruction fine-tuning data. The first type, represented by the LLaVA dataset (Liu et al. 2023b), is manually curated and combined with language-only GPT-4 (OpenAI 2023b) for multimodal models. It includes 150K training samples, subdivided into simple dialogue (57,669 samples), detailed description (23,240 samples), and complex reasoning vision-language data (76,803 samples). This dataset spans various facets of multimodal dialogue, including category recognition, counting, action recognition, and scene recognition. The detailed descriptions demand careful image observation and comprehensive detailing, while the complex reasoning tasks require deep inference and external knowledge integration. The second type of data is multimodal instruction fine-tuning data derived from publicly available image-text datasets. Specifically, we used OKVQA (Marino et al. 2019) and A-OKVQA (Schwenk et al. 2022) datasets, as utilized in InstructBLIP (Dai et al. 2023), for VIGC training. These datasets, necessitating extensive external knowledge, are ideal for assessing the VIGC's capabilities.

Inference Data. Following the VIGC network training, we generated fine-tuning data for multimodal instruction using image datasets. We employed two distinct datasets, COCO (Lin et al. 2014) and Objects365 (Shao et al. 2019), to evaluate VIGC's effectiveness in handling data within the same or different image domains. The COCO dataset serves as the foundation for the construction of the LLaVA, OKVQA, and A-OKVQA datasets. It's crucial to emphasize that during the data generation phase, we intentionally omitted any images that were previously included in the test set to ensure the fairness and effectiveness of the evaluation.

Implementation Details

During the training phase of VIGC, we utilize the MiniGPT-4 (Zhu et al. 2023b) first-stage pre-trained model as the source of initial parameters. This ensures that the initial model does not incorporate additional instruction fine-tuning data for training, thereby preserving the fairness of downstream task validation. This model encompasses the ViT-G/14 from EVA-CLIP (Fang et al. 2023), the Q-Former (Li et al. 2023b), and a linear projection layer. The language models employed are Vicuna7B and Vicuna13B (Chiang et al. 2023). It is noteworthy that, as illustrated in Figure 1, our Q-Former is designed to receive either Instruction or Question text simultaneously, which is crucial for the iterative correction in VIC. Therefore, we utilize the Q-Former from BLIP2-FlanT5$_{\mathrm{XXL}}$ (Li et al. 2023b) as the initial parameters for the Q-Former. We designate this network model as MiniGPT-4+. During the training process, only the parameters of the Q-Former and the linear projection layer are subjected to fine-tuning, while the parameters of the language and visual models remain constant. The training is conducted throughout 10 epochs, with the model's performance being validated after each epoch. The model that demonstrates the best performance is subsequently selected for data generation.

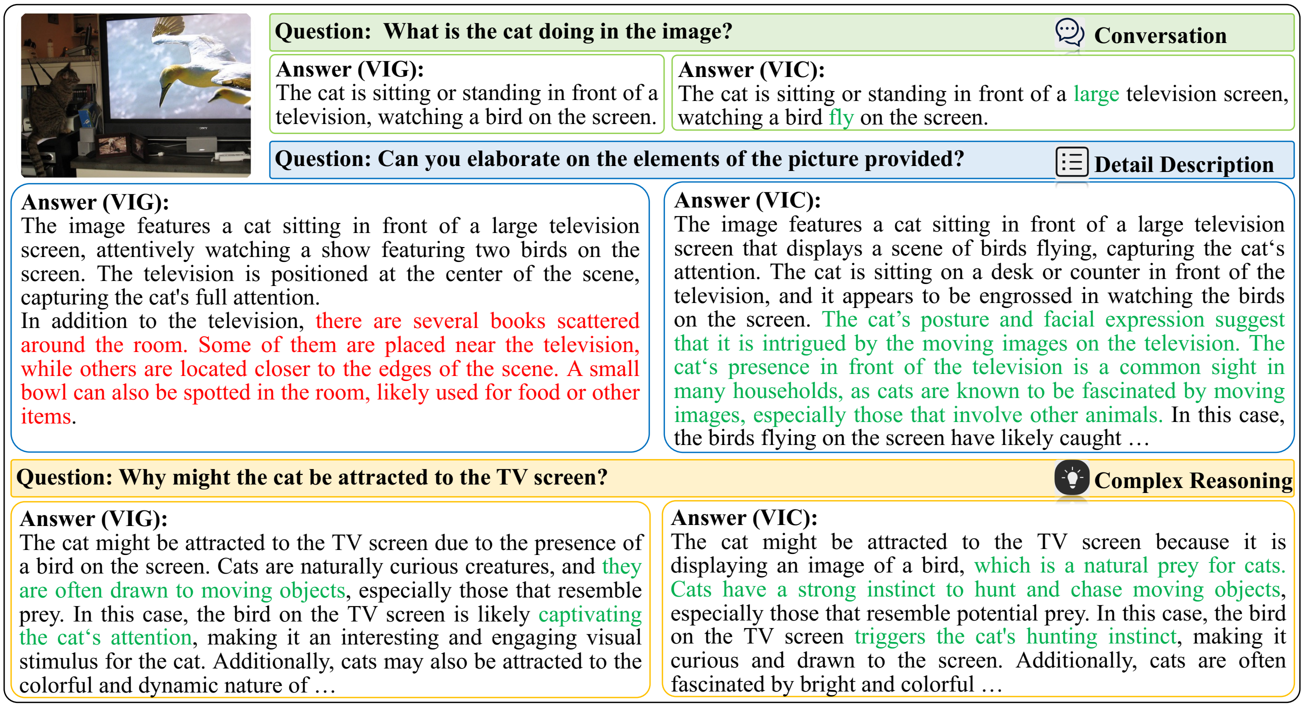
In terms of batch sizes, we utilize 64 for both 7B and 13B models. The entire training process, executed on 8 A100 (80GB) GPUs, completes in approximately 10 hours.

LLaVA Data and Evaluation

Dataset Analysis. In the pursuit of generating a more diverse set of LLaVA-like data, the VIGC model is trained using a combination of LLaVA-150K data and three types of instruction templates. During the inference phase, we utilized images from the COCO 2017 training set, intentionally excluding those already included in the LLaVA dataset. This resulted in the selection of a total of 36,781 initial images, which served as the foundation for instruction data generation; we refer to this data as **coco-extra**, which serves as the default supplementary data used for model training during evaluation.

Based on the aforementioned data, the VIG network generates diverse initial questions and answers. Subsequently, the VIC network refines the outputs by taking the questions and the existing answers as inputs through the Iterative Q-Former (IQF) operation, thus generating higher-quality responses. Figure 4 illustrates the three categories of data generated via the VIGC process:

- Conversation: The questions are typically specific, eliciting concise and clear responses.
- Detail Description: The questions are relatively fixed and revolve around describing the image's content. This re-



Question: What is the cat doing in the image?

Answer (VIG):
The cat is sitting or standing in front of a television, watching a bird on the screen.

Answer (VIC):
The cat is sitting or standing in front of a large television screen, watching a bird fly on the screen.

Question: Can you elaborate on the elements of the picture provided?

Answer (VIG):
The image features a cat sitting in front of a large television screen, attentively watching a show featuring two birds on the screen. The television is positioned at the center of the scene, capturing the cat's full attention.
In addition to the television, there are several books scattered around the room. Some of them are placed near the television, while others are located closer to the edges of the scene. A small bowl can also be spotted in the room, likely used for food or other items.

Answer (VIC):
The image features a cat sitting in front of a large television screen that displays a scene of birds flying, capturing the cat's attention. The cat is sitting on a desk or counter in front of the television, and it appears to be engrossed in watching the birds on the screen. The cat's posture and facial expression suggest that it is intrigued by the moving images on the television. The cat's presence in front of the television is a common sight in many households, as cats are known to be fascinated by moving images, especially those that involve other animals. In this case, the birds flying on the screen have likely caught …

Question: Why might the cat be attracted to the TV screen?

Answer (VIG):
The cat might be attracted to the TV screen due to the presence of a bird on the screen. Cats are naturally curious creatures, and they are often drawn to moving objects, especially those that resemble prey. In this case, the bird on the TV screen is likely captivating the cat's attention, making it an interesting and engaging visual stimulus for the cat. Additionally, cats may also be attracted to the colorful and dynamic nature of …

Answer (VIC):
The cat might be attracted to the TV screen because it is displaying an image of a bird, which is a natural prey for cats. Cats have a strong instinct to hunt and chase moving objects, especially those that resemble potential prey. In this case, the bird on the TV screen triggers the cat's hunting instinct, making it curious and drawn to the screen. Additionally, cats are often fascinated by bright and colorful …

Figure 4: Generated instructions based on the proposed VIGC.

Model	Metrics			
	Conv	Detail	Complex	All
LLaVA-7B (Baseline)	75.1	75.4	92.3	81.0
add 36K Conv (VIG)	80.9	76.1	92.6	83.3
add 36K Conv (VIC)	83.9	76.9	90.9	**84.0**
add 36K Detail (VIG)	80.2	72.7	90.9	81.4
add 36K Detail (VIC)	83.3	80.6	93.1	**85.8**
add 36K Complex (VIG)	81.4	75.6	90.5	82.6
add 36K Complex (VIC)	80.2	76.2	93.2	**83.3**
replace 10K Conv	78.2	76.5	91.6	82.1
replace 10K Detail	75.8	79.8	91.2	82.2
replace 10K Complex	77.5	77.8	92.8	82.8
replace Combined	78.3	76.6	92.4	82.5

Table 1: Comparative evaluation of VIGC data addition vs. replacement in model training on the LLaVA evaluation

quires the model to clearly observe all targets within the image. It is observed that the detailed descriptions generated directly from VIG are fraught with numerous illusions. However, after the application of VIC, these illusory phenomena have significantly diminished.

- Complex Reasoning: The posed questions necessitate the integration of external knowledge and the application of sophisticated logical reasoning skills.

Overall, the quality of the question-answer pairs autonomously generated by the model has exceeded our initial expectations. We posit that this rich new knowledge inherently resides within the language model itself, and we have merely employed multimodal instruction fine-tuning to distill this knowledge onto new multimodal data.

Dataset Evaluation. Based on the generated data, we conducted detailed ablation experiments on LLaVA-7B to verify the performance improvement of the model after training with the generated data. The evaluation method used here is the quantitative evaluation proposed by LLaVA, where GPT-4 assesses the quality of the model's responses to given evaluation questions, which can be understood as relative scores compared to GPT-4. LLaVA provides 30 test images, each containing three types of questions, for 90 questions.

Table 1 presents the results of augmenting the original LLaVA-150K dataset with three types of generated data, followed by fine-tuning the LLaVA first-stage model with instructions. Including instruction data directly generated from VIG during the training phase has proven to be beneficial. We observed a marginal improvement when adding detailed description data generated by VIG, which can be attributed to the severe illusions present in this data. In contrast, the incorporation of conversation and complex reasoning data has led to appreciable performance gains.

Further refining the data using VIC and then training the model with the augmented conversation data, detailed description data, and complex reasoning data resulted in additional improvements. The performance metrics have reached 84.0%, 85.8%, and 83.3%, respectively. These results underscore the critical role of VIC in eliminating hallucinations, thereby enhancing the model's overall performance. Simul-

Method	MMBench							LLaVA			
	LR	AR	RR	FP-S	FP-C	CP	Overall	Conv	Detail	Complex	All
MiniGPT-4+	10.0	31.3	7.83	18.9	13.1	43.0	24.4	83.5	77.8	92.4	84.7
MiniGPT-4+ w/ coco	11.7	27.8	19.1	27.9	11.0	44.3	**27.5**(↑ 3.1)	84.1	84.1	92.7	**87.0**(↑ 2.3)

Table 2: Performance of MiniGPT-4+ models on MMBench and LLaVA-eval datasets. MMBench metric include logic reasoning (LR), attribute reasoning (AR), relation reasoning (RR), fine-grained perception at instance-level (FP-S), fine-grained perception at cross-instance (FP-C), and coarse perception (CP).

Model	Conv	Detail	Complex	All
LLaVA-7B	75.1	75.4	92.3	81.0
w/ coco	83.3	80.6	93.1	**85.8**(↑ 4.8)
w/ objects365	86.8	77.6	90.9	85.2
LLaVA-13B*	82.7	76.6	94.8	84.8
w/ coco	88.9	77.4	93.5	**86.8**(↑ 2.0)

Table 3: Relative scores for different settings w.r.t. GPT-4 (language-only) on LLaVA-eval Dataset. The results for LLaVA-13B are reproduced from (Liu et al. 2023b).

Model	OKVQA	A-OKVQA
PaLM-E (Driess et al. 2023)	66.1	-
PromptCap (Hu et al. 2022)	60.4	56.3
MiniGPT-4+ w/o VIGC	59.1	58.3
MiniGPT-4+ w/ VIGC	**59.8**(↑ 0.7)	**58.9**(↑ 0.6)
InstructBLIP w/o VIGC	63.1	62.5
InstructBLIP w/ VIGC	**63.8** (↑ 0.7)	**64.1** (↑ 1.6)

Table 4: Results of finetuning MiniGPT-4+ and InstructBLIP on OKVQA and A-OKVQA dataset.

taneously, to validate the superiority of the VIGC-generated dataset over the LLaVA dataset, we conducted an experiment where we randomly replaced 10,000 instances from each type of data, as well as a complete replacement of all three types of data. The experimental results indicated that, under the condition of constant data volume, the performance of the model trained on a mixture of the LLaVA dataset and the VIGC dataset surpasses that of the model trained solely on the LLaVA dataset.

Table 3 presents experiments conducted on different datasets and models of varying sizes, substantiating that the use of generated data from different domains, such as Objects365 and COCO, can still lead to remarkable performance improvements. This offers a novel solution for enhancing the performance of cross-domain tasks. We also conducted experiments on LLaVA-13B, proving that performance enhancement can be achieved on larger models.

We also evaluated the performance of the VIGC model on MMBench, LLaVA (as shown in Table 2) and further finetuned the VIGC model based on 36K COCO data generated by VIGC. We discovered that following this self-iterative training process, the model performance improved on both MMBench and LLaVA. This promising capability of self-enhancement through iterative training is a subject we plan to continue exploring in our future research.

OK-VQA Dataset and Evaluation

To further assess the quality of the data generated by the VIGC model, we conducted training and evaluation on the OKVQA dataset, which requires external knowledge. Specifically, we trained the VIGC network using the OKVQA dataset and corresponding instruction templates. Subsequently, we generated additional instruction fine-tuning data based on VIGC on COCO. Ultimately, we fine-tuned InstructBLIP based on OKVQA and the generated data. We found that despite InstructBLIP already utilizing a substantial amount of data in the instruction fine-tuning phase, the use of additional generated data for downstream task fine-tuning still enhanced the model's performance on specific datasets. We performed the same experimental validation on A-OKVQA.

The experimental results are presented in Table 4. It can be seen that the performance of the InstructBLIP model, when fine-tuned with the addition of generated data, outperforms the model only fine-tuned with original data. There were improvements of 0.7% and 1.6% on OKVQA and A-OKVQA, respectively, achieving state-of-the-art results for models of this scale on both datasets. BUsing the MiniGPT-4+ pre-training model, we arrived at similar conclusions. This demonstrates that generated data can effectively enhance downstream fine-tuning performance, a finding that holds significant value for domains where data acquisition is challenging.

Conclusion

We introduce the Visual Instruction Generation and Correction (VIGC) framework for generating high-quality vision-language instruction data. Using VIGC, we produced diverse, validated multimodal instruction data on COCO and Objects365 datasets. The framework provides an efficient means for enhancing instruction tuning data. While VIGC significantly reduces model hallucination, some instances persist, necessitating further exploration into multimodal hallucination solutions.

Acknowledgments

This project was supported by National Key R&D Program of China (No. 2022ZD0160101) and Shanghai Artificial Intelligence Laboratory.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *NeurIPS*.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2305.06500*.

Driess, D.; Xia, F.; Sajjadi, M. S.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*.

He, C.; Jin, Z.; Xu, C.; Qiu, J.; Wang, B.; Li, W.; Yan, H.; Wang, J.; and Lin, D. 2023. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models. *arXiv preprint arXiv:2308.10755*.

He, C.; Li, W.; Jin, Z.; Wang, B.; Xu, C.; and Lin, D. 2022. OpenDataLab: Empowering General Artificial Intelligence with Open Datasets. https://opendatalab.com. Accessed: 2023-12-22.

Hu, Y.; Hua, H.; Yang, Z.; Shi, W.; Smith, N. A.; and Luo, J. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2023. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. *arXiv preprint arXiv:2311.17911*.

Iyer, S.; Lin, X. V.; Pasunuru, R.; Mihaylov, T.; Simig, D.; Yu, P.; Shuster, K.; Wang, T.; Liu, Q.; Koura, P. S.; et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.

Li, B.; Zhang, Y.; Chen, L.; Wang, J.; Pu, F.; Yang, J.; Li, C.; and Liu, Z. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, Y.; Duan, N.; Zhou, B.; Chu, X.; Ouyang, W.; Wang, X.; and Zhou, M. 2018. Visual question generation as dual task of visual question answering. In *CVPR*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*. Springer.

Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; and Wang, L. 2023a. Aligning Large Multi-Modal Model with Robust Instruction Tuning. *arXiv preprint arXiv:2306.14565*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*.

Mostafazadeh, N.; Misra, I.; Devlin, J.; Mitchell, M.; He, X.; and Vanderwende, L. 2016. Generating natural questions about an image. In *ACL*.

OpenAI. 2023a. ChatGPT. https://openai.com/blog/chatgpt. Accessed: 2023-12-22.

OpenAI. 2023b. GPT-4 Technical Report. arXiv:2303.08774.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Peng, B.; Li, C.; He, P.; Galley, M.; and Gao, J. 2023a. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023b. Kosmos-2: Grounding Multimodal Large Language Models to the World. *arXiv preprint arXiv:2306.14824*.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*.

Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vedd, N.; Wang, Z.; Rei, M.; Miao, Y.; and Specia, L. 2021. Guiding visual question generation. In *NAACL*.

Wang, Y.; Mishra, S.; Alipoormolabashi, P.; Kordi, Y.; Mirzaei, A.; Arunkumar, A.; Ashok, A.; Dhanasekaran, A. S.; Naik, A.; Stap, D.; et al. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *EMNLP*.

You, H.; Sun, R.; Wang, Z.; Chen, L.; Wang, G.; Ayyubi, H. A.; Chang, K.-W.; and Chang, S.-F. 2023. IdealGPT: Iteratively Decomposing Vision and Language Reasoning via Large Language Models. *arXiv preprint arXiv:2305.14985*.

Zhang, P.; Wang, X. D. B.; Cao, Y.; Xu, C.; Ouyang, L.; Zhao, Z.; Ding, S.; Zhang, S.; Duan, H.; Yan, H.; et al. 2023. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhao, Z.; Wang, B.; Ouyang, L.; Dong, X.; Wang, J.; and He, C. 2023. Beyond Hallucinations: Enhancing LVLMs through Hallucination-Aware Direct Preference Optimization. *arXiv preprint arXiv:2311.16839*.

Zhu, D.; Chen, J.; Haydarov, K.; Shen, X.; Zhang, W.; and Elhoseiny, M. 2023a. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023b. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.